Does data citation aid provenance? — An update from ESIP

Mark Parsons[1], Ruth Duerr[2], Nancy Hoebelheinrich[3], Sophie Hou[4], Matt Mayernik[4], and Hampapuram Ramapriyan[5]

[1]Rensselaer Polytechnic Institute
[2]Ronin Institute
[3]Knowledge Motifs LLC
[4]National Center for Atmospheric Research
[5]Science Systems and Applications, Inc./NASA Goddard Space Fight Center

November 26, 2022

Abstract

Formal data citation is a growing practice increasingly required by scientific journals. Roughly a decade ago, the Federation of Earth Science Information Partners (ESIP) began developing formal guidelines for data citation including acknowledgement of authors and archives and careful use of persistent identifiers (PIDs). Many Earth science data centers now provide formal citation text and PIDs for their data sets, typically a Digital Object Identifier (DOI). A central purpose of data citation (amongst many) is to aid scientific reproducibility through direct, unambiguous reference to the precise data used in a particular study, i.e., to aid provenance tracking. How has that worked in practice? ESIP is now in the process of revising and updating their guidelines and seeks to ensure that data citation meets its stated purpose. This presentation explores whether and how formal citation and the the use of PIDs for data sets has improved the tracking of data provenance. For example, is there is some commonality in the nature and granularity of objects that are assigned PIDs? We review how the guidelines are being revised to further enhance the transparency and reusability of data.

Does data citation aid provenance? — An update from ESIP

Mark A. Parsons[1], Ruth Duerr[2], Nancy Hoebelheinrich[3], Sophie Hou[4], Matt Mayernik[4], Hampapuram Ramapriyan[5], and the ESIP Preservation and Stewardship Committee
[1] Rensselaer Polytechnic Institute, [2] Ronin Institute, [3] Knowledge Motifs LLC, [4] National Center for Atmospheric Research, [5] Science Systems and Applications, Inc./NASA Goddard Space Fight Center
@ESIPfed

"A data citation is a reference to data for the purpose of credit attribution and facilitation of access to the data." (CODATA-ICSTI 2013).

This aligns with the first Recommendation of the W3C Provenance Incubator Group (2010) that there "should be a standard way to represent at a minimum three basic provenance entities:

- a handle (URI) to refer to an object (resource)
- a person/entity that the object is attributed to
- a processing step done by a person/entity to an object to create a new object"

So does citation aid provenance? Yes, but only a little bit.

Citation was designed for people to identify and credit scholarly resources. Now with persistent identifiers we seek to accomplish more machine readability, access, and interchange.

We find that the existing model is only partially adaptable to the networked representation of the research enterprise that we view as necessary for full understanding of provenance.

Motivation and Background: Evolving the ESIP Data Citation Guidelines

In January 2012, the Earth Science Information Partners (ESIP) formally endorsed guidelines to help repositories develop data citations (ESIP Stewardship Committee 2012). The guidelines have been widely adopted by Earth science data centers and will now be recommended in in the author guidelines of most Earth science journals (Stall et al., 2018).

There has been much community discussion about the particulars of data citation since the ESIP Guidelines were endorsed, including the Joint Declaration of Data Citation Principles (Data Citation Synthesis Group 2014), Research Data Alliance Recommendations on dynamic data citation (Rauber et al 2016) and link exchange between literature and data (Scholix) (Burton et al. 2017), and emergent guidance on software citation (Smith et al. 2016; Katz & Hong 2018).

Given these developments, the new imperative coming from both journals and data centers, and the lack of simple instructions on how to construct and resolve a basic citation for either data or software, the ESIP Data Stewardship Committee is revising their guidelines. We seek to ensure we meet the basic requirements of the CODATA-ICSTI definition above and the Joint Declaration of Data Citation Principles, including aspects of provenance,

We begin with a very basic use case. So basic that we hope it applies for data, software, and other research objects.

A "simple" use case

"Researcher cites a relatively static, whole, scientific resource (data, software, or other research object) in the scholarly literature."

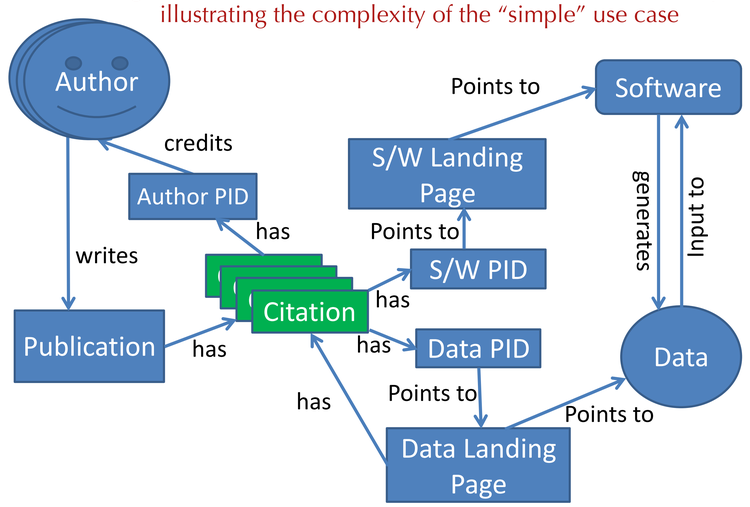
Repositories, therefore, need to:

A. Provide the researcher a recommended citation

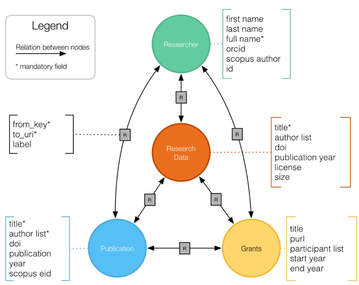
B. Provide a mechanism to resolve that citation

> "The meaning of meaning is relationship."
> — Marshall McLuhan

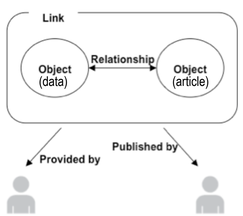
High-level model of citation-related entities and relationships illustrating the complexity of the "simple" use case



Two closely related but contrasting RDA recommendations develop different approaches on how to relate research objects



Data Description Registry Interoperability (DDRI) graph (courtesy Amir Aryani)



Scholix Framework (Burton et. al 2017)

Finding	Implication
Citation only addresses a narrow use case, which is only a small part of provenance.	Don't expect too much from basic citation. It captures just a few of many relationships.
Software citation is ad hoc or nonexistent (Li et al. 2018). Data citation is not much better, even for data journals (Stuart 2017).	Use existing systems like CrossRef and DataCite, but explore other possibilities.
Current systems are largely geared around existing scholarly registries and are not well inter-networked. This constrains what can actually be "cited". It's not for us to decide what is important, what is registered (Parsons & Fox 2018), or what is referenced.	Can we have a generic research object citation protocol? Examining new DONA Foundation Digital Object Interface Protocol (2018).
Resource type is very contextual, but it is necessary for provenance.	Unclear how or whether to implement in a citation context.
Granularity and data packaging are inconsistent, and there is uneven consensus on how best to capture and represent parent-child relationships. Similarly, the nature of the relationship between article and data or software is inconsistent (see lower figure).	Probably best to keep it vague for now — allow late semantic binding and apply the robustness principle. Postel's Law: Be conservative in what you do, be liberal in what you accept from others.
With the growth of PIDs there is a corresponding growth of multiple PIDs pointing to the same (or very similar) object, plus a proliferation of non-canonical URIs that complicate the landscape, especially around credit. Metadata schemas and data systems are still adapting to the use of identifiers.	Recognize the different timescales for human adaptation vs. changes in embedded digital infrastructure.

Next Steps

1. Update current guidelines to address the simple use case.
   a. Provide guidance for more complex use cases such as dynamic data and multi-sourced products, but still within the context of literature. The new guidelines will be a more dynamic document.
   b. Clarify purpose, content, and structure of landing pages. Identify the role of schema.org (pulling and indexing instead of pushing and registering). What about direct content access?
2. Expand the concept to consider broader use cases of reference and credit. Rethink the primacy of citation. It may not fit into the network view. Does the approach really work for all types of research objects, including physical objects?

Get involved!

Join the **ESIP Preservation and Stewardship Committee**: http://wiki.esipfed.org/index.php/Preservation_and_Stewardship or the new **Citation Cluster** https://lists.esipfed.org/mailman/listinfo/esip-citationguidelines

Burton, A, A Aryani, H Koers, P Manghi, S La Bruzzo, M Stocker, M Diepenbroek, U Schindler, and M Fenner. 2017. "The Scholix Framework for Interoperability in Data-Literature Information Exchange." *D-Lib Magazine 23* (1/2): https://doi.org/10.1045/january2017-burton.

CODATA-ICSTI Task Group on Data Citation Standards and Practices. 2013. "Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data." *Data Science Journal 12* (0): CIDCR1–CIDCR75. https://doi.org/10.2481/dsj.OSOM13-043.

Data Citation Synthesis Group. 2014. *Joint Declaration of Data Citation Principles*. Martone, M (ed.). Force11. https://doi.org10.25490/a97f-egyk.

ESIP Stewardship Committee. 2012. *Data Citation Guidelines for Data Providers and Archives*. Parsons, MA et al. (eds.). ESIP Federation. https://doi.org/10.7269/P34F1NNJ.

Katz, DS, and N. Chue Hong. 2018. "Software Citation in Theory and Practice." *Arxiv* preprint https://arxiv.org/pdf/1807.08149.pdf.

Li, K, P-Y Chen, and E Yan. 2018. "Challenges of measuring the impact of software: an examination of the lme4 R package." *Arxiv* Preprint https://arxiv.org/pdf/1811.11270.pdf.

Parsons, MA., and PA. Fox. 2018. "Power and Persistent Identifiers." *International Data Week 2018* https://doi.org/0.5281/zenodo.1495322.

Rauber, A, A Asmi, D van Uytvanck, and S Pröll. 2016. "Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use." *Bulletin of IEEE Technical Committee on Digital Libraries 12* (1): https://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf.

Smith, AM., DS Katz, KE Niemeyer, and Software Citation Working Group FORCE11. 2016. "Software citation principles." *PeerJ Computer Science* 2 e86. https://doi.org/10.7717/peerj-cs.86.

Stall, S, L Yarmey, et al.. 2018. "Advancing FAIR Data in Earth, Space, and Environmental Science." *Eos 99* https://doi.org/10.1029/2018eo109301.

Stuart, D. 2017. "Data bibliometrics: metrics before norms." *Online Info Review 41* (3): 428--35. https://doi.org/10.1108/oir-01-2017-0008.

W3C Provenance Incubator Group. 2010. *Provenance XG Final Report: W3C Incubator Group Report 08 December 2010*. Gil, Y, J Cheney, P Groth, O Hartig, L Moreau, and P Pinheiro da Silva (eds.). http://www.w3.org/2005/Incubator/prov/XGR-prov/.